UIDS: A Multilingual Document Summarization Framework Based on Summary Diversity and Hierarchical Topics

Lei Li, Yazhao Zhang, Junqi Chi, Zuying Huang

Center for Intelligence Science and Technology, School of Computer
Beijing University of Posts and Telecommunications, Beijing, P.R. China
{leili,yazhao}@bupt.edu.cn,709722796@qq.com,hpwthzy@126.com

Abstract. In this paper, we put forward UIDS, a new high-performing extensible framework for extractive MultiLingual Document Summarization. Our approach looks on a document in a multilingual corpus as an item sequence set, in which each sentence is an item sequence and each item is the minimal semantic unit. Then we formalize the extractive summary as summary diversity sampling problem that considers topic diversity and redundancy at the same time. The topic diversity is reflected using hierarchical topic models, the redundancy is reflected using similarity and the summary diversity is enhanced using Determinantal Point Processes. We then illustrate how this method encompasses a framework that is amenable to compute summaries for MultiLingual Single- and Multi-documents. Experiments on the MultiLing summarization task datasets demonstrate the effectiveness of our approach.

Keywords: Multilingual Document Summarization · Summary Diversity · Determinantal Point Processes

1 Introduction

With the development of information communication technology, a large number of electronic documents are created on the Internet. Under this circumstance, it is an enormous challenge for users to find concise and relevant information, especially in the cross-language case. This motivates the development of methods to compute summaries for multiple languages.

Document summarization can generally be classified into extractive and abstractive ones. An abstractive summary can be seen as a reproduction of the original document in a new way. However, an extractive summary is generated by selecting a few relevant sentences from the original document.

The task on which we focus in this paper is extractive MultiLingual Document Summarization(MDS). The goal of MDS is to compute a summary for every document or document cluster in the multilingual corpus. The language of the summary is consistent with the language of the corresponding source text. And the summary length depends on the compressing rate which is mostly decided by users.

In this paper, our goal is to systematically explore an amenable unsupervised language independent sampling framework to automatically calculate diverse summaries (Unsupervised language Independent Diverse Summary method, UIDS) for multilingual corpus. According to our observation, a good summary should be diverse in latent semantics, including topic diversity (quality) and redundancy. The three main contributions of our work are:

1. We focus on the issue of Summary Diversity that contains *Topic Diversity* and *Redundancy* of extractive summarization.
2. We enhance the summary diversity of Multilingual documents using Determinantal Point Processes and multiple features including hierarchical topic models.
3. We propose an efficient extensible language independent framework to solve the Multilingual Document Summarization task.

The rest of this paper is organized as follows: Section 2 presents related works. Section 3 defines the summary diversity property and MultiLingual Document Summarization. Section 4 presents the system framework of UIDS based on hierarchical Topic Model(HTM) and Determinantal Point Processes(DPPs). A set of experiments are implemented on the MultiLingual datasets, described in Section 5. In section 6, we conclude the paper with some pointers to future research directions.

2 Related Work

An extractive summary is often viewed as a machine learning problem: selecting a subset of sentences from a given document[9]. And several effective methods have been proposed to solve MDS problem.

In the method of [15], features were categorized as surface, content, relevance and event. They combined these four features to select sentences to compute a summary. In SIGDial 2015, UJF-Grenoble [2], CCS [7], EXB [20], NTNU [14] and UA-SLAI [21] all presented their systems on MSS task. CCS [7] used a term weighting method called OCCAMS[8] to compute the weight of each sentence and then choose the top few sentences to construct the summary.

Previous work also proved the effectiveness of Topic Models. [4] first proposed Latent Dirichlet Allocation(LDA) and used it into summarization. To relax the assumption that topic count of LDA is known and fixed, [3] extended LDA to exploit the hierarchical tree structure of topics, called hierarchical Latent Dirichlet Allocation, hLDA in short. It organizes the topics into a tree-like structure and supposes the topic count could grow with the dataset automatically. [6] provides a multi-document summarization based on supervised hLDA and obtains competitive results. [13] used hLDA based multiple feature combination method to compute MultiLingual Multi-Document Summarization.

The diversity property has been researched in many areas other than summarization, such as biodiversity, Shannon's diversity index and so on. [18] considered

the diversity for summarization. They put forward a contrastive theme summarization based on hLDA and Structured Determinantal Point Processes(SDPPs). They use topic probability of word, under viewpoint to calculate qualities. But unfortunately, their focus on solving the sentiment diversity. Our method differs from the above methods by emphasizing the latent semantic diversity property implicated in summarization, which will be proved to be a novel solution for multilingual summarization.

3 Motivation and Formalization

3.1 Summary Diversity

As presented in classic topic models, a document can be represented using multiple latent topics. And each sentence belongs to a topic according to a certain probability. The consensus amongst the researchers of topic model is that each topic has some kind of latent semantic information implicated in articles. Besides, diversity is an instance of being composed of differing elements or qualities. Based on this, we believe that a good summary should be diverse in latent semantic level. That is to say a diverse summary should satisfy the following two requirements:

1. **Topic Diversity:** the summary should contain those important topics hidden in the document and each topic cannot be similar to others.
2. **Redundancy:** the summary should not contain two or more sentences that describe the same topic or aspect.

In general, our goal is to compute summaries with diversity. In extractive case, a summary may be a subset of sentences.

3.2 MultiLingual Single Document Summarization

Given a *document* D from a multilingual corpus, we can represent it using a sentence set $\{s_1, ..., s_n\}$, where n is the total number of sentences. Every *sentence* is a sequence of words $s_i = \{w_1, ..., w_m\}$. It is a big challenge to understand the specific meaning of every word, especially in the case of multiple languages. For English, a word is a sequence of letters. However, for Chinese, a word is one or more Chinese characters. In this paper, the "word" will be collectively referred to as an *item*. It represents the minimal semantic unit in a sentence that depends on its language.

Following the topic modeling customs, we define a *topic* in a document D to be a probability distribution over items. Different from "flat" topic model, we assume that the topics in D are organized as a tree-like hierarchy and every node is a topic. Every sentence s_i is assigned to a path c_j from the root node to a leaf node.

Given a document represented using items and topics, we define the *quality* of a sentence and the *similarity* between sentences. The *quality* of s_i determines

the degree of its reflection on topics. However, the *similarity* determines the redundancy between sentences.

Finally, we define extractive MultiLingual Document Summary. Given a document $D = \{s_1, \ldots, s_n\}$ represented using items. The purpose of extractive summarization for MDS is to sample a subset $D^{'} \subset D$ that gives consideration to both *quality* and *similarity*, thus generating a diverse summary covering more important information of the original document. And the sentence sampling method must be language independent.

4 *UIDS* Framework

In this section we describe our *UIDS* framework. It is presented as a *framework* rather than a singular approach because a number of the implementation details can very depending on the purpose of the task(Multi-document or Single document). And it provides a generic structure for building more specialized systems. Fig. 1 shows the framework of our unsupervised language independent diverse summarization system (UIDS) based on Summary Diversity and DPPs. The line of dashes in Fig. 1 means that it can be modified according to specific tasks.

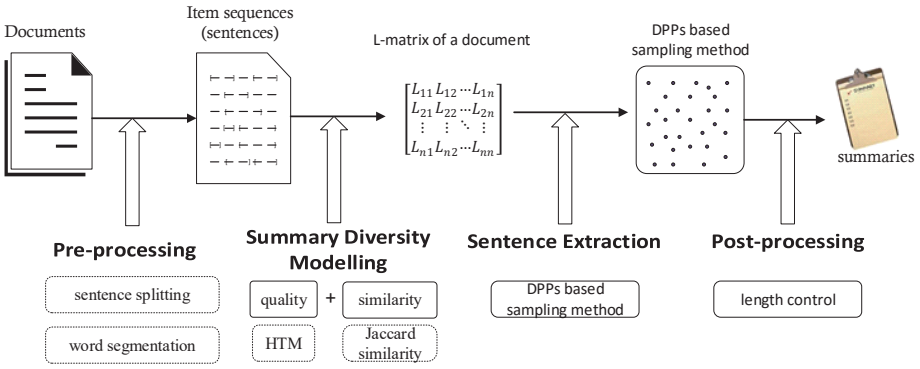

Fig. 1. Framework of UIDS system

4.1 Pre-processing

To deal with the linguistic difference and convert the documents into item sequences, we use the following steps to pre-process the original documents. (1) spilt document into sentences using period for every document; (2) use Rosette API [19] to tokenize the document of some languages, such as Chinese, Thai and Japanese; (3) calculate TF-IDF value of each item in every document; (4) model every document using hLDA [3].

4.2 Summary Diversity Modelling

Given the document and hierarchical topics calculated in 4.1, we estimate the quality (topic diversity) of each sentence using hierarchical topic information.

hierarchical Topic Model (HTM): hLDA constructs a document to a tree-like structure. Based on this, we adapt the method that [13] proposed to use the information of hLDA for MSS task. HTM represents the level score of item calculated according to the topics in hLDA. We give the item higher score whose level distribution is more close to that of items in golden summary. We use Equation (1) to calculate it.

$$q_{HTM} = \sum_{i=1}^{m} (\alpha_i T_i + Freq_i) \tag{1}$$

where T_i represents the distribution score of $item_i$ calculated by hLDA, α_i is the pre-defined weight of T_i according to our former experiments, $Freq_i$ is the frequency of $item_i$ in current hLDA node.

4.3 Determinantal Point Processes for Diverse Summary Extraction

Arisen in quantum physics and random matrix theory, DPPs are elegant probabilistic models of global, negative correlations [16]. In this paper, we will focus on discrete DPPs and follow the definition of Kulesza [1].

A point process P on a discrete set $Y = \{x_1, x_2, ..., x_n\}$ is a probability measure on 2^Y, the set of all subsets of Y. A Determinantal Point Process is a point process with a positive semidefinite matrix K, which is indexed by the elements of Y. That is the i-th row of K corresponds to the i-th element in Y. Thus the definition of DPP is:

Definition 1. *When Y is a random subset drawing according to point process P, we have, for every $A \subseteq Y$,*

$$P(A \subseteq \mathbf{Y}) = \det(K_A)$$

where $\det(K_A) = |K_{ij}|_{i,j \in A}$ *and we adopt* $\det(K_\Phi) = 1$.

As for K contains all information needed to compute the probability of any subset A being included in Y, we call it as *kernel matrix*. In order to model real data, we use L-ensemble [5] to construct DPPs. Thus we have

$$P_L(\mathbf{Y} = Y) = \frac{\det(L_Y)}{\det(L + I)}.$$

where L is a positive semidefinite matrix, $\det(L_Y) = |L_{ij}|_{i,j \in Y}$, I is the $N \times N$ identity matrix.

Using this representation, the entries of kernel L can be written as

$$L_{ij} = q_i \phi_i^\top \phi_j q_j$$

where $q_i \in R^+$ measures the *quality* of an element i, and $\phi_i^\top \phi_j$ is often regarded as a whole that measures the *similarity* between element i and element j.

To compute a summary, we adopt sampling algorithm proposed by Kulesza and Taskar to sample a sentence subset from a document, shown in Table 1. As we can see that the sampling algorithm is language independent.

Input: HTM, S, D, max_len.
$\rightarrow$ $quality\_vec = \omega_0$*HTM
$\rightarrow$ matrix_l= $quality\_vec * S * quality\_vec^T$
$\rightarrow$ $(\mathbf{v}_n, \lambda_n)$ =eigen_decompose(matrix_l)
$\rightarrow$ $J = \emptyset$
$\rightarrow$ **for** $n = 1, 2, \ldots, N$ **do**
$\rightarrow\rightarrow$ $J = J \cup \{n\}$ with prob. $\frac{\lambda_n}{\lambda_n + 1}$
$\rightarrow$ $V = \{\mathbf{v}_n\}_{n \in J}$
$\rightarrow$ $Y = \emptyset$
$\rightarrow$ **while** $|V| > 0$ and $len(Y) <$ max_len **do**
$\rightarrow\rightarrow$ Select i from Y with $Pr(i) = \frac{1}{|V|} \sum_{v \in V} (\mathbf{v}^\top \mathbf{e}_i)^2$
$\rightarrow\rightarrow$ $Y = Y \cup D[i]$
$\rightarrow\rightarrow$ $V = V_\perp$, an orthonormal basis for the subspace of V orthogonal to e_i
Output: summary Y

Table 1. DPPs sampling method for diverse summary extraction, where S is the similarity matrix, D is the document, ω_0 is the parameter of sentence quality calculated using HTM

Besides, we use JACCARD to measure the sentence similarity (Redundancy), shown in Equation(2)

$$r_{ij} = \frac{|s_i \cap s_j|}{|s_i \cup s_j|} \tag{2}$$

Thus each element L_{ij} can be calculated as follows.

$$L_{ij} = q_i r_{ij} q_j \tag{3}$$

Finally, we will truncate summary to human summary length and remove multiple white spaces to get our final summaries in the post-processing step.

5 Experiments

To test the capability of UIDS summarization framework, we also propose two contrast systems shown below.

Combined Features (CF): except *HTM* introduced in Section 4.2, we also propose four features to calculate the *score* of each sentence, then we select the top few sentences to construct the summary.

1. *Sentence Position (SP)*: Here we use Equation (4) to calculate the score of sentence position.

$$q_{SP} = \frac{n - i + 1}{n} \tag{4}$$

where n is the total sentence number of the document, i represents $i - th$ sentence in the document.

2. *Title Similarity (TS)*: Title Similarity is the cosine similarity of the sentence and the document title. We use Equation (5) to calculate it.

$$q_{TS} = \frac{tf_{s_i} \times tf_{s_{title}}}{|tf_{s_i}||tf_{s_{title}}|} \tag{5}$$

where s_{title} and s_i represent the title and a sentence respectively.

3. *Sentence Length (SL)*: We define the sentence length as *item* number in s_i. Then we use Gaussian distribution to normalize the score, shown in Equation (6).

$$q_{SL} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(L_i - \mu)^2}{2\sigma^2}} \tag{6}$$

where L_i is the length of sentence i, μ is the average length and σ^2 is the variance of lengths.

4. *Sentence Coverage (SC)*: If an item appears in many sentences, then the sentence containing the item has a high probability of being selected as summary. Based on this, we can calculate SC using Equation (7)

$$q_{SC} = \frac{\sum_{i=1}^{|s|} \frac{num_s(item_i)}{n}}{|s|} \tag{7}$$

where $num_s(item_i)$ is the sentence number that contains $item_i$, $|s|$ is the sentence length.
Given the five features (HTM, SP, TS, SL, SC), the score calculation method is shown below:

$$score_i = \sum_{k=1}^{5} \varphi_k q_{ki} \tag{8}$$

where $\varphi_k \in \{0, 1\}$ is the combination proportion of corresponding features. q_{ki} represents the k-th quality feature of s_i.

Graphical Model: graphical model can obtain the summary in an unsupervised way, and it is also corpus and language independent. As a comparison, we use the LexRank algorithm [17] to obtain the score of sentence s_i. We then calculate the sentence order on the basis of sentence score. Finally, we select sentences one by one until the length of selected sentences exceeds the length limit. The method is shown in Table. 2.

Input: sim_{ij}, Φ, max_inter, d
for $i = 1, 2, \ldots, N$ **do**:
$\rightarrow$ **for** $j = 1, 2, \ldots, N$ **do**:
$\rightarrow\rightarrow$ **if** $sim_{ij} > \Phi$ **then**:
$\rightarrow\rightarrow\rightarrow$ $Edge_{ij} \leftarrow sim_{ij}$
$\rightarrow\rightarrow$ **else**
$\rightarrow\rightarrow\rightarrow$ $Edge_{ij} \leftarrow 0$
$\rightarrow$ $LR_i \leftarrow \frac{1-d}{N}$
$iter = 1$
while $iter < max\_iter$ **do**
$\rightarrow$ **for** $i = 1, 2, \ldots, N$ **do**
$\rightarrow\rightarrow$ $LR^{'} = LR_i$
$\rightarrow\rightarrow$ **for** $j = 1, 2, \ldots, N$ **do**
$\rightarrow\rightarrow\rightarrow$ $LR_i \leftarrow \frac{1-d}{N} + d\frac{sim_{ji}}{\sum_{k=1}^{N} sim_{jk}} LR_j^{'}$
Output: LR

Table 2. Graphical Model Method for MSS, where sim_{ij} is the similarity of s_i, s_j, Φ is the edge threshold, max_inter is the max iteration times, d is the damping coefficient

5.1 Multilingual Single Document Summarization

We follow the MSS-2017 task[12] at MultiLing-2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres implemented within EACL 2017 (15th Conference of the European Chapter of the Association for Computational Linguistics), so as to measure the performance of our framework and methods.

The testing dataset contains 30 featured Wikipedia articles from each of 24 languages. All documents in testing dataset are provided by MSS-2017 task and formatted in both an XML format and raw text. The documents are in UTF-8 without mark-ups and images. And the target summary length is the same as the character length of the human summary.

Following existing models, we set predefined parameters (ω_0) of UIDS to the sum of the quality of every sentence(UIDS-HTM). In order to test the performance of HTM, we also use SP (UIDS-SP) and TS (UIDS-TS) to replace it to calculate the quality of sentences, and then construct matrix L. Besides CF and GM, there are two teams **SWAP** and **TeamMD** participated in the MSS task and they proposed nine summary methods. The organizer of MSS task has also provided three MSS systems for comparisons. **Oracle** uses the combinatorial covering algorithm in [8] by selecting sentences from its body text to cover the items in the human summary. Thus it can be considered as an upper bound approximation for every document. **Lead** just cuts the original document to summary length. **IWB** uses the structure of the document sections to extract sentences from sections and subsections. Table 3 shows the identifier of every method.

Table. 4 gives the ROUGE-2 F scores of 12 systems for 24 languages. The first column in both tables contain the ISO code for each language [10]. As

Method Name	Identifier	Method Name	Identifier
IWB	B1	**Combined Feature**	CF
Lead	B2	**Graphical Model**	GM
Oracle	B3	**TeamMD-1**	T1
UIDS-HTM	S1	**TeamMD-2**	T2
UIDS-SP	S2	**TeamMD-3**	T3
UIDS-TS	S3	**TeamMD-4**	T4

Table 3. System names and its supporting languages

shown in Table. 4, S3 (UIDS-TS) performs much better than S2 (UIDS-SP) in 16 languages. And the phenomenon may due to the fact that titles of Wikipedia articles are highly summative. But anyway this illustrates that the performances of both TS and SP are deficient in stability. Fortunately, S1 (UIDS-HTM) performs much better than S2 and S3 in almost every language. This indicates that HTM based latent topic modeling method is more useful for latent topic diversity and sentence diversity sampling.

For more observation, Oracle ranks first in 24 languages with no doubt as it is the upper bound approximation of every document. With no consideration to Oracle, CF and GM methods give the worst performance in most languages. S1 performs better than other methods in 11 languages, and ranks top 3 in every language. This fully demonstrates the effectiveness of our framework.

We also notice that documents in az, bs, jv, li, lv, mr, tt, uk are much harder for summarization than others. Because MultiLing-2017 does not provide any training data for those languages. Our approach UIDS ranks top 2 in those languages and outperforms the B1 (IWB) baseline in 4 kinds of languages. This demonstrates the robustness of our framework. And this phenomenon may due to the fact that in the generation of a good summarization, the latent central idea based sentence quality measurement and summary diversity property can partly replace the role played by training data.

5.2 MultiLingual Multi-document Summarization

The MultiLing hold annual workshops and adjoining competitions to encourage research in Multilingual Document Summarization. We present the results on the testing dataset of MultiLing-2015 MMS task (MMS-2015). MMS-2015 data is made up of from 10 to 15 sets of 10 news articles for each of 10 languages.

We report our results using the n-gram matching metric ROUGE following the setting of [11], shown in Table. 5. *UIDS* is compared to several systems that participated in the MMS task. There are 7 systems in MultiLing2015 MMS task involving a human summarization. Despite using only HTM described in Section. 4.2, *UIDS* performs best in Chinese, Czech and Hebrew. However, in Arabic, Greek, Hindi, the performance of our system is bad. Especially for Hindi, our results are far less than others. We need more experiments later to explore the reasons for this situation. But on the whole, our system is still competitive.

lang	B1	B2	B3	S1	S2	S3	CF	GM	T1	T2	T3	T4
ar	5.982	4.916	8.958	**5.395**	3.98	4.61	3.03	2.999	4.523	4.008	3.887	3.706
az	4.79	3.928	11.216	**5.143**	4.422	4.143	2.636	1.64	4.249	3.886	4.151	3.875
bg	6.224	4.365	12.712	**6.009**	4.918	5.677	3.154	3.296	5.205	5.637	4.601	4.46
bs	4.407	3.485	9.053	**3.943**	3.03	3.403	2.74	3.617	3.192	3.815	2.881	3.446
cs	5.877	4.019	11.811	**5.442**	4.354	4.823	3.064	4.665	4.39	4.723	4.137	4.203
en	13.318	10.931	20.623	**13.985**	11.891	12.794	7.751	13.437	11.93	12.438	10.453	10.293
eo	8.331	6.916	12.682	**7.81**	6.934	7.554	4.868	3.17	6.729	6.973	5.662	5.823
fi	8.642	2.241	10.397	**5.148**	3.529	4.032	3.421	0	3.932	3.826	3.75	3.575
hr	4.475	3.196	8.704	**4.691**	3.006	3.05	3.1	2.973	3.123	3.714	3.47	3.126
id	9.206	6.437	14.354	**8.586**	7.682	7.969	5.153	7.973	8.362	7.599	7.082	6.821
jv	6.962	6.287	9.447	**8.309**	5.068	5.957	4.74	5.17	5.495	5.682	4.478	5.23
ko	2.792	2.387	6.763	2.385	**2.719**	2.144	1.805	1.602	1.838	1.954	2.013	2.079
li	4.365	3.405	6.103	**4.713**	3.286	3.059	3.357	3.831	4.596	4.504	3.744	3.344
lv	6.03	3.964	10.426	**4.594**	3.413	4.515	1.669	3.715	3.579	4.164	3.661	3.287
mr	18.84	17.787	27.584	**19.521**	17.24	16.986	10.654	18.411	19.427	18.469	16.725	16.547
ms	8.308	6.003	10.581	**7.033**	5.863	6.404	5.089	3.23	6.577	5.905	5.248	4.57
pl	6.302	4.03	11.981	**6.104**	4.462	5.673	4.02	3.934	4.685	4.63	4.641	4.233
pt	11.237	7.708	16.939	**10.453**	8.304	9.651	4.685	7.92	9.974	10.07	7.979	7.642
ro	8.751	6.113	15.281	**8.482**	7.955	7.496	4.092	7.541	7.794	7.961	6.936	7.096
sk	3.244	2.609	6.968	2.77	**2.844**	2.262	2.118	2.325	2.687	2.083	2.598	1.947
tr	6.928	6.072	12.409	**7.085**	6.765	6.316	4.27	6.108	5.419	4.797	5.834	5.429
tt	2.614	2.614	5.335	**2.787**	2.002	2.481	1.539	2.705	1.92	2.095	2.105	2.205
uk	3.462	1.855	6.782	**2.907**	1.829	1.644	1.14	1.253	2.041	1.649	1.456	1.769
zh	18.292	16.821	20.289	14.632	14.713	**15.128**	9.709	11.817	16.307	17.654	16.245	16.03

Table 4. ROUGE-2 results of all languages

lang	ours	cist	esi	giau	human	mms3	occams	wbu
Arabic	0.10028	0.13729	0.18379	0.19055	0.49773	0.19645	0.21181	0.23783
Chinese	**0.20145**	0.04935	0.13445	-	0.55843	0.20097	0.06799	0.15947
Czech	**0.20914**	0.12686	0.14192	0.19723	0.46779	0.17239	0.18494	0.20409
English	0.12240	0.10529	0.16764	0.14924	0.46165	0.16849	0.17961	0.19215
French	0.19343	0.12148	0.19442	0.17381	0.4919	0.21797	0.20702	0.2484
Greek	0.12375	0.1206	0.12624	0.1755	0.44346	0.14829	0.15989	0.16623
Hebrew	**0.09428**	0.04968	0.09124	0.09653	0.37887	0.08192	0.07677	0.09337
Hindi	0.11192	0.24051	0.18841	0.25874	0.62603	0.27152	0.25877	0.28622
Romanian	0.12281	0.09476	0.1271	-	0.69859	0.14403	0.16013	0.18904
Spanish	0.23137	0.13581	0.22059	-	0.52141	0.24963	0.2444	0.28102

Table 5. ROUGE-2 results of MMS-2015 task, where - represents that the system does not participate the language

6 Conclusion and Future Work

In this paper, we have systematically explored an amenable language independent sampling framework to automatically calculate diverse summaries for mul-

tilingual corpus. We formalize MDS using summary diversity property, which includes topic diversity and redundancy. We have also detailed a basic implementation of MMS and MSS system that outperformed some of the most advanced methods described in the literature.

In the present description of *UIDS*, we use five classical features including HTM, SP, SL, SC, TS to calculate topic diversity and Jaccard similarity to model redundancy. This might not be feasible for a large-scale summary creation method. For this, we will pay more attention to DPPs. Then we will extend our system to social data summarization and so on.

Acknowledgements This work was supported by the National Social Science Foundation of China under Grant 16ZDA055; National Natural Science Foundation of China under Grant 91546121, 71231002 and 61202247; EU FP7 IRSES MobileCloud Project 612212; the 111 Project of China under Grant B08004; Engineering Research Center of Information Networks, Ministry of Education; the project of Beijing Institute of Science and Technology Information; the project of CapInfo Company Limited.

References

1. Alex, K., Ben, T.: Determinantal point processes for machine learning. arXiv preprint arXiv:1207.6083 (2012)
2. Balikas, G., Amini, M.R.: The participation of ujf-grenoble team at multiling 2015 (2015)
3. Blei, D.M., Griffiths, T.L., Jordan, M.I.: The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. Journal of the ACM (JACM) 57(2), 7 (2010)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research 3(Jan), 993–1022 (2003)
5. Borodin, A.: Determinantal point processes (2009)
6. Celikyilmaz, A., Hakkani-Tur, D.: A hybrid hierarchical model for multi-document summarization. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. pp. 815–824. Association for Computational Linguistics (2010)
7. Conroy, J.M., Davis, S.T., Kubina, J.: Preprocessing and termweights in multilingual summarization (2015)
8. Davis, S.T., Conroy, J.M., Schlesinger, J.D.: Occams–an optimal combinatorial covering algorithm for multi-document summarization. In: Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on. pp. 454–463. IEEE (2012)
9. Gambhir, M., Gupta, V.: Recent automatic text summarization techniques: a survey. Artificial Intelligence Review 47(1), 1–66 (2017)
10. George Giannakopoulos1, John Conroy2, J.K.P.A.: Multiling 2017 overview (2017)
11. Giannakopoulos, G., Kubina, J., Conroy, J.M., Steinberger, J., Favre, B., Kabadjov, M.A., Kruschwitz, U., Poesio, M.: Multiling 2015: Multilingual summarization of single and multi-documents, on-line fora, and call-center conversations. In: SIGDIAL Conference. pp. 270–274 (2015)

12. Giannakopoulos, G., Lloret, E., Conroy, M.J., Steinberger, J., Litvak, M., Rankel, P., Favre, B.: Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres, chap. Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres. Association for Computational Linguistics (2017), http://aclweb.org/anthology/W17-1000
13. Huang, T., Li, L., Zhang, Y.: Multilingual multi-document summarization based on multiple feature combination (2016)
14. Hung, H.T., Shih, K.W., Chen, B.: The ntnu summarization system at multiling 2015 (2015)
15. Kam-Fai, W., Mingli, W., Wenjie, L.: Extractive summarization using supervised and semi-supervised learning. In: Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1. pp. 985–992. Association for Computational Linguistics (2008)
16. Matérn, B.: Stochastic previous models and their application to some problems in forest surveys and other sampling investigations. Medd. Statens Skogsforskningsinstitut 49(5) (1960)
17. Mihalcea, R., Tarau, P.: Textrank: Bringing order into texts. In: Proceedings of EMNLP 2004. pp. 404–411 (2004)
18. Ren, Z., de Rijke, M.: Summarizing contrastive themes via hierarchical non-parametric processes. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 93–102. ACM (2015)
19. Technology, B.: Rosette base linguistics. https://www.rosette.com/function/tokenization/ (2016)
20. Thomas, S., Beutenmüller, C., de la Puente, X., Remus, R., Bordag, S.: Exb text summarizer. In: 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue. p. 260 (2015)
21. Vicente, M., Alcón, O., Lloret, E.: The university of alicante at multiling 2015: approach, results and further insights. In: 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue. p. 250 (2015)